ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ Ι

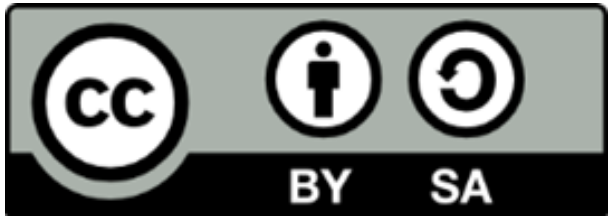
Ενότητα 10:
Εισαγωγικά περί κανονικοποίησης
Συναρτησιακές εξαρτήσεις – BCNF

Ευαγγελίδης Γεώργιος
Τμήμα Εφαρμοσμένης Πληροφορικής

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.



Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Μακεδονίας» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Μέρος 1

Εισαγωγικά περί κανονικοποίησης

Πως προκύπτει ένα σχήμα;

- **Top-down** με σχεδίαση ενός διαγράμματος Οντοτήτων-Συσχετίσεων και μετατροπή αυτού σε ένα σύνολο πινάκων (σχέσεων) [δες ενότητες 2 και 3].
- **Bottom-up** με αυτοματοποιημένο αλγοριθμικό τρόπο (χρήση της **θεωρίας κανονικοποίησης**).

Παράδειγμα με CDBASE (1)

Επίπεδος (flat) πίνακας, παρόμοιος με λογιστικό φύλλο:

TRACK(cd_title, year, company_name, track_position, song_title, performer_name, composer_name, lyricist_name)

Πρόβλημα: Πως ξεχωρίζω cd με τον ίδιο τίτλο, τραγούδια με τον ίδιο τίτλο, εταιρίες / ερμηνευτές / συνθέτες / στιχουργούς με το ίδιο όνομα;

Λύση: χρήση κάποιου έξτρα πεδίου που παίρνει μοναδικές τιμές για κάθε ένα από αυτά τα αντικείμενα (οντότητες;)

Παράδειγμα με CDBASE (2)

- Μια βελτιωμένη εκδοχή:

TRACK(cd_id, cd_title, year, company_id, company_name, track_position, song_id, song_title, performer_id, performer_name, composer_id, composer_name, lyricist_id, lyricist_name)

Παράδειγμα με CDBASE (3)

- Αλλά, όπως ήδη είδαμε έχουμε **περιττή επανάληψη πληροφορίας** καθώς και όλα τα αρνητικά επακόλουθα (στις ενημερώσεις και στις διαγραφές).
- Πόσες εγγραφές χρειαζόμαστε για ένα track που έχει 4 ερμηνευτές, 3 στιχουργούς και 2 συνθέτες;
- Στο παραπάνω σενάριο (δηλαδή για την καταγραφή της πληροφορίας ενός μόνο track), πόσες φορές δηλώνουμε κάποια πληροφορία για το cd και την εταιρία; - για κάθε ερμηνευτή / συνθέτη / στιχουργό;
- Πρόκειται για ένα **κακό** σχήμα.

Παράδειγμα με CDBASE (4)

Τώρα “μπαίνουν στο παιχνίδι” και οι **επιχειρησιακοί κανόνες** που προκύπτουν κατά τη φάση της **ανάλυσης απαιτήσεων**:

- Μπορεί να υπάρχουν πολλαπλές εκδοχές (ηχογραφήσεις) ενός τραγουδιού;
- Μπορεί το ίδιο τραγούδι να εμφανίζεται στο ίδιο cd είτε με την ίδια είτε με διαφορετική ηχογράφηση;
- Μπορεί ένα τραγούδι να έχει πολλούς συνθέτες ή στιχουργούς;
- Μπορεί μια ηχογράφηση να έχει πολλούς ερμηνευτές;

Ένα καλύτερο σχήμα

SONG(song_id, song_title)

COMPOSER(composer_id, composer_name)

LYRICIST(lyricist_id, lyricist_name)

SONG_COMP(song_id, composer_id)

SONG_LYR(song_id, lyricist_id)

RECORDING(rec_id, song_id)

PERFORMER(performer_id, performer_name)

REC_PERF(rec_id, performer_id)

TRACK(cd_id, track_pos, rec_id)

CD(cd_id, cd_title, year, company_id)

COMPANY(company_id, company_name)

Σχεδίαση με κανονικοποίηση (1)

Πως φτάνω όμως στο προηγούμενο “καλό” σχήμα;

- Ξεκινώ από έναν **universal** πίνακα.
- Διασπώ σε **μικρότερους καλύτερους** πίνακες.
- Η όλη διαδικασία γίνεται με αυτοματοποιημένο τρόπο (**αλγόριθμος**).

Σχεδίαση με κανονικοποίηση (2)

- Για να δουλέψει ο αλγόριθμος χρειάζεται επιπλέον πληροφορία εκτός της **universal σχέσης**.
- Είναι πληροφορία που σχετίζεται με τους επιχειρησιακούς κανόνες, δηλαδή με τις **ιδιότητες που έχουν τα δεδομένα**.
- Ο αλγόριθμος παράγει ένα σύνολο πινάκων που ικανοποιούν μια **κανονική μορφή** (normal form).

Σχεδίαση με κανονικοποίηση (3)

- **Συναρτησιακές εξαρτήσεις ==> Boyce-Codd NF**
- **Εξαρτήσεις πολλαπλών τιμών ==> 4NF**

Συναρτησιακή εξάρτηση = Functional dependency

Εξάρτηση πολλαπλών τιμών = Multivalued dependency

- Στα επόμενα θα χρησιμοποιούμε τις συντομογραφίες FD και MVD.

Συναρτησιακές εξαρτήσεις

TRACK(cd_id, cd_title, year, rec_id, track_pos)

- Κακός σχεδιασμός.
- Έχουμε την FD: cd_id → cd_title, year

BCNF: αν A → B τότε το A πρέπει να είναι κλειδί

- Άρα διασπώ τον TRACK σε:

CD(cd_id, cd_title, year)

TRACK(cd_id, rec_id, track_pos)

Εξαρτήσεις πολλαπλών τιμών

SONG(song_id, composer_id, lyricist_id)

- Κακός σχεδιασμός παρόλο που είναι σε BCNF.
- Πρακτικά, δεν υπάρχει καμία FD στον Song.
- Όμως έχω την MVD: song_id —≫ composer_id

4NF: αν A —≫ B τότε το A πρέπει να είναι κλειδί

- Άρα διασπώ τον SONG σε:

SONG_COMP(song_id, composer_id)

SONG_LYR(song_id, lyricist_id)

Μέρος 2

Συναρτησιακές εξαρτήσεις – BCNF

Ορισμός FD

A → B

- Όταν σε δυο διαφορετικές εγγραφές ενός πίνακα τα Α έχουν την ίδια τιμή τότε και τα Β έχουν την ίδια τιμή.
- Αλλιώς, το Α προσδιορίζει το Β.
- Γενίκευση:

A1, A2, ... An → B1, B2, ... Bm ή X → Y

- Παράδειγμα: **cd_id → cd_title, year**

FD και κλειδί

- Έχω έναν πίνακα R χωρίς διπλότυπα.
- Έστω ότι **A, B → όλα-τα-πεδία-του-R.**
- Αυτός όμως είναι **ορισμός του κλειδιού.**
- Στα παρακάτω χρησιμοποιούμε τους συμβολισμούς A, B, C για πεδία και X, Y, Z για σύνολα πεδίων.

Είδη FD

- Τετριμμένες:

X → Y και Y ⊆ X

- Μη-τετριμμένες:

X → Y και Y ⊈ X

- Πλήρως μη-τετριμμένες

X → Y και X ∩ Y = ∅

Μας ενδιαφέρουν μόνο FD του τελευταίου είδους.

Κανόνες για FD (1)

Κανόνας διαχωρισμού:

Αν **X → B1, B2, ... Bm** τότε

X → B1

X → B2

...

X → Bm

Το ανάποδο δεν ισχύει!

Κανόνες για FD (2)

Κανόνας συνένωσης:

Αν

X → B1

X → B2

...

X → Bm τότε

X → B1, B2, ... Bm

Κανόνες για FD (3)

Τετριμμένοι κανόνες:

X → **Y** τότε **X** → **X** ∪ **Y**

X → **Y** τότε **X** → **X** ∩ **Y**

Μεταβατικός κανόνας:

X → **Y** και **Y** → **Z** τότε **X** → **Z**

Εγκλεισμός πεδίων

- Έχω έναν πίνακα R, ένα σύνολο από FDs, και ένα σύνολο πεδίων X του R.
- Εγκλεισμός του X, ή $\mathbf{X^+}$**,** είναι το σύνολο όλων των πεδίων Y του R για τα οποία ισχύει $\mathbf{X \rightarrow Y}$.
- Αλγόριθμος; Ξεκινώ θέτοντας $\mathbf{X^+ = X}$ και μετά για κάθε FD $\mathbf{Y \rightarrow Z}$ **με** $\mathbf{Y \subseteq X}$ προσθέτω το Z στο X^+. Συνεχίζω μέχρι να μην μπορεί να προστεθεί κάτι στο X^+.

Παράδειγμα εγκλεισμού

RECORDING(song_id, song_title, rec_id, performer_id, performer_name) και ισχύουν

song_id → song_title

performer_id → performer_name

rec_id → song_id

Τότε **{performer_id, rec_id}$^+$ = {όλα τα πεδία}**

Εγκλεισμός και κλειδιά

- Πότε ξέρουμε αν **X είναι κλειδί του R**;
- Αν **X^+ ισούται με το σύνολο των πεδίων του R**!
- Άρα αν έχουμε έναν πίνακα R και ένα σύνολο από FDs, πως βρίσκουμε τα κλειδιά του R;
- Αλγόριθμος: δοκιμάζουμε αν A^+ είναι κλειδί για κάθε πεδίο A του R. Αν δεν βρούμε κανένα κλειδί, δοκιμάζουμε με ζεύγη πεδίων, κοκ.

Καθορίζοντας FDs για πίνακα (1)

- Αν έχω δυο σύνολα S1 και S2 από FDs για έναν πίνακα R, τότε το **S2 προκύπτει από το S1** αν κάθε εγγραφή του R που ικανοποιεί τις FDs του S1 ικανοποιεί και τις FDs του S2.
- Παράδειγμα: το S2 = {rec_id → song_title} προκύπτει από το S1 = {rec_id → song_id, song_id → song_title}
- Πως βρίσκω αν X → Y προκύπτει από ένα σύνολο S από FDs; Υπολογίζω το X^+ βάσει του S και ελέγχω αν το Y ανήκει στο X^+.

Καθορίζοντας FDs για πίνακα (2)

- Ποιες FDs θέλουμε για έναν πίνακα R;
- **Το ελάχιστο σύνολο από πλήρως μη-τετριμμένες FDs ώστε όλες οι FDs που ισχύουν για τον πίνακα R να προκύπτουν από αυτό το σύνολο.**
- Θα το δούμε στην πράξη!

Διάσπαση πίνακα

Έχουμε μια **καλή διάσπαση** όταν

ο αρχικός πίνακας **R(X)**

διασπάται στους πίνακες **R1(Y)** και **R2(Z)**

ώστε **Y** ∪ **Z = X** και **R1 |x| R2 = R**

Κανονικοποίηση με διάσπαση

- Ξεκινάμε από έναν universal πίνακα και ιδιότητες των δεδομένων (FDs).
- Ο αλγόριθμος διασπά βάσει των FDs.
- Στο τέλος μετά από μια σειρά καλών διασπάσεων έχουμε ένα σύνολο καλών πινάκων (που είναι σε BCNF).

Ορισμός BCNF

- Ένας πίνακας R με FDs είναι σε BCNF αν

Για κάθε X → B, το X είναι κλειδί

Παράδειγμα BCNF (1)

R(song_id, song_title, rec_id, performer_id, performer_name)

και ισχύουν

song_id → song_title

performer_id → performer_name

rec_id → song_id

Το ελάχιστο κλειδί του R είναι το **{performer_id, rec_id}.**

Βάσει των τριών FD ο R δεν είναι σε BCNF.

Διασπώ...

Παράδειγμα BCNF (2)

R1(song_id, song_title) είναι σε BCNF

R2(song_id, rec_id, performer_id, performer_name)
συνεχίζω τη διάσπαση

R21(performer_id, performer_name) είναι σε BCNF

R22(song_id, rec_id, performer_id)

συνεχίζω τη διάσπαση

R221(rec_id, song_id) είναι σε BCNF

R222(rec_id, performer_id) είναι σε BCNF

Αλγόριθμος διάσπασης BCNF

- **Input**: πίνακας R και FDs του R
- **Output**: διάσπαση του R σε σύνολο πινάκων σε BCNF
- Υπολόγισε με τη βοήθεια των FDs τα κλειδιά του R.
- Επανέλαβε μέχρι να είναι όλοι οι πίνακες σε BCNF:
- Πάρε έναν πίνακα **R'** ο οποίος εξαιτίας μιας **X → Y** να **μην είναι σε BCNF**.
- Διάσπασε τον **R'** σε **R1(X, Y)** και **R2(X, υπόλοιπα-πεδία)**.
- Υπολόγισε τις FDs των R1 και R2.
- Υπολόγισε τα κλειδιά των R1 και R2.

Γιατί οι BCNF πίνακες είναι καλοί

- Δεν έχουν προβλήματα **πλεονασμού**.
- Μπορούν να **αναδημιουργήσουν τους αρχικούς πίνακες** με σύζευξη.

Πλήρες παράδειγμα (1)

T(cd_id, cd_title, year, company_id, company_name, track_position, track_duration, rec_id, rec_duration, song_id, song_title, performer_id, performer_name, composer_id, composer_name, lyricist_id, lyricist_name)

1. song_id → song_title

2. composer_id → composer_name

3. lyricist_id → lyricist_name

4. rec_id → song_id, rec_duration

5. performer_id → performer_name

6. cd_id, track_position → rec_id, track_duration

7. cd_id → cd_title, year, company_id

8. company_id → company_name

Ελάχιστο κλειδί = {cd_id, track_position, performer_id, composer_id, lyricist_id)

Πλήρες παράδειγμα (2)

Ο Τ δεν είναι σε BCNF. Διασπώ με την FD

1. song_id → song_title

T1(song_id, song_title) BCNF

T2(cd_id, cd_title, year, company_id, company_name, track_position, track_duration, rec_id, rec_duration, song_id, performer_id, performer_name, composer_id, composer_name, lyricist_id, lyricist_name)

Κλειδί = {cd_id, track_position, performer_id, composer_id, lyricist_id}, FDs = {2, 3, 4, 5, 6, 7, 8}.

Πλήρες παράδειγμα (3)

Ο T2 δεν είναι σε BCNF. Διασπώ με τις FDs **2, 3, 5, 7, 8**:

T3(composer_id, composer_name) BCNF

T4(lyricist_id, lyricist_name) BCNF

T5(performer_id, performer_name) BCNF

T6(cd_id, cd_title, year, company_id) BCNF

T7(company_id, company_name) BCNF

T8(cd_id, track_position, track_duration, rec_id, rec_duration, song_id, performer_id, composer_id, lyricist_id)

Κλειδί = {cd_id, track_position, performer_id, composer_id, lyricist_id}FDs = {4, 6}.

Πλήρες παράδειγμα (4)

Ο Τ8 δεν είναι σε BCNF. Διασπώ με την

4. rec_id → song_id, rec_duration

T9(<u>rec_id</u>, song_id, rec_duration) BCNF

T10(cd_id, track_position, track_duration, rec_id, performer_id, composer_id, lyricist_id)

Κλειδί = {cd_id, track_position, performer_id, composer_id, lyricist_id}, FDs = {6}.

Πλήρες παράδειγμα (5)

Ο T10 δεν είναι σε BCNF. Διασπώ με την

6. cd_id, track_position → rec_id, track_duration

T11(cd_id, track_position, rec_id, track_duration) BCNF

T12(cd_id, track_position, performer_id, composer_id, lyricist_id) BCNF

Κλειδί = {cd_id, track_position, performer_id, composer_id, lyricist_id}, FDs = {}.

Πλήρες παράδειγμα (6)

Τελικό BCNF Σχήμα:

SONG(song_id, song_title)

COMPOSER(composer_id, composer_name)

LYRICIST(lyricist_id, lyricist_name)

PERFORMER(performer_id, performer_name)

CD(cd_id, cd_title, year, company_id)

COMPANY(company_id, company_name)

RECORDING(rec_id, song_id, rec_duration)

TRACK(cd_id, track_position, rec_id, track_duration)

TRACK_CONTRIB(cd_id, track_position, performer_id, composer_id, lyricist_id)

ανοικτάμαθήματα
opencourses

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

Τέλος Ενότητας

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης